(19) JAPANESE PATENT OFFICE

PATENT ABSTRACTS OF JAPAN

(11) Publication number: 09311802 A

(43) Date of publication of application: 02.12.97

(51) Int. Cl G06F 12/00
G06F 12/00
G06F 11/34
G06F 15/00
G06F 17/30

(21) Application number: 08149784

(22) Date of filing: 22.05.96

(71) Applicant: MATSUSHITA ELECTRIC IND CO LTD

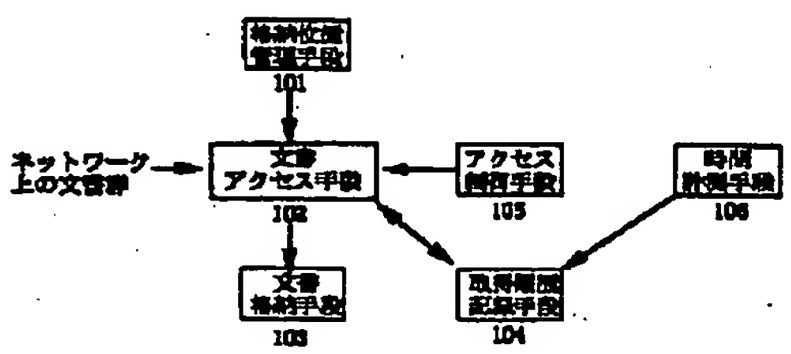
(72) Inventor: UENO TAKESHI
NOGUCHI YOSHIHIRO
SATO MITSUHIRO
ISHIKAWA MIKITO

(54) DOCUMENT GATHERING DEVICE

(57) Abstract:

PROBLEM TO BE SOLVED: To provide a document gathering device for automatically and efficiently gathering always newest document information at the time of gathering document in a network.

SOLUTION: By the start from an access control means 105, a document access means 102 gathers a desired document from a document group on a network by using a storing position managing means 101 managing the correspondence of the document position on the network and the document name, a document storing means stores it, and a time measuring means 106 records time at that time in an obtained history recording means 104. At the time of gathering document after then, only documents updated after the time recorded in the recording means 104 are gathered.



(19)日本国特許庁（ＪＰ）　(12)公開特許公報（A）　(11)特許出願公開番号

特開平9－311802

(43)公開日　平成9年(1997)12月2日

(51)Int.Cl.⁶		識別記号	庁内整理番号	FI		技術表示箇所
G06F	12/00	517		G06F	12/00	517
		545				545M
	11/34				11/34	C
	15/00	310			15/00	310U
	17/30				15/401	340B

審査請求　未請求　請求項の数6　ＦＤ　(全 10 頁)

(21)出願番号　特願平8－149784

(22)出願日　平成8年(1996)5月22日

(71)出願人　000005821
松下電器産業株式会社
大阪府門真市大字門真1006番地

(72)発明者　上野　剛
大阪府門真市大字門真1006番地　松下電器産業株式会社内

(72)発明者　野口　喜祥
大阪府門真市大字門真1006番地　松下電器産業株式会社内

(72)発明者　佐藤　光弘
大阪府門真市大字門真1006番地　松下電器産業株式会社内

(74)代理人　弁理士　役　昌明　(外2名)

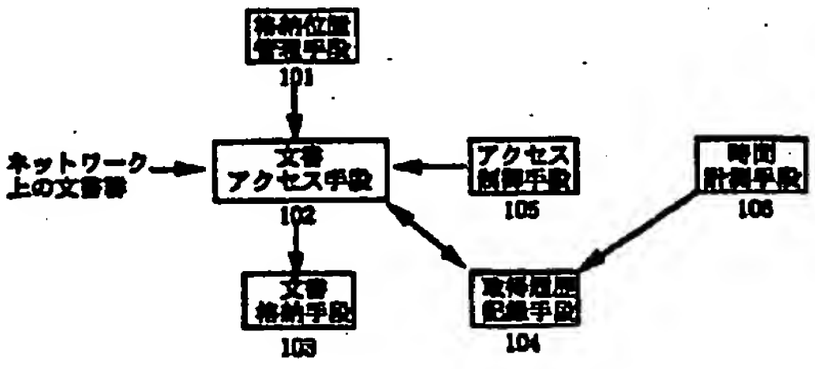
最終頁に続く

(54)【発明の名称】　文書収集装置

(57)【要約】

【課題】　ネットワーク上にある文書を収集する時に、常に最新の文書情報を自動的にかつ効率良く収集するための文書収集装置を提供すること。

【解決手段】　アクセス制御手段105からの起動により、ネットワーク上の文書位置と文書名の対応を管理する格納位置管理手段101を用いて文書アクセス手段102がネットワーク上の文書群から目的の文書を収集し、文書格納手段103に格納し、時間計測手段106により、その時の時刻を取得履歴記録手段106に記録し、その後の文書収集で前記取得履歴記録手段104に記録された時刻以降に更新された文書のみを収集する。



1

【特許請求の範囲】

【請求項1】　ネットワーク上に分散格納されそれぞれ別個に更新される文書群のデータを該ネットワークに接続された特定の計算機内に自動的に収集する文書収集装置において、文書収集をする文書アクセス手段と，それを駆動するアクセス制御手段をもうけ、前回収集した時点以降に更新された文書のみを収集することを特徴とする文書収集装置。

【請求項2】　文書が格納されているネットワーク上の位置と文書の対応を管理する格納位置管理手段と、ネットワーク上の文書の内容を計算機内に読み込む文書アクセス手段と、読み込んだ文書の内容を記憶する文書格納手段と、時間を計測する時間計測手段と、前記文書を読み込んだ日時を前記時間計測手段からの時刻で記録する取得履歴記録手段と、前記文書を最後に取得した取得日時とネットワーク上での前記文書の更新日時を比較し前回収集した時点以降に更新された文書のみを収集するように前記文書アクセス手段を制御するアクセス制御手段から構成されることを特徴とする文書収集装置。

【請求項3】　文書とともに得られた更新日時を更新履歴として記録する更新履歴記録手段を更に備え、前記更新履歴記録手段に記録された文書毎の複数の更新日時の記録により、前記アクセス制御手段が前記文書の平均的な更新周期を計算し前記文書を前記更新周期毎かつ予測される更新日時の直後に取得するように前記時間計測手段を利用して前記文書アクセス手段を制御することを特徴とする請求項2記載の文書収集装置。

【請求項4】　前記更新履歴記録手段に記録された文書毎の複数の更新日時の記録により更新パターンを抽出する更新パターン抽出手段と、その更新パターンを記録する更新パターン記録手段とを更に備え、前記アクセス制御手段が前記文書を前記更新パターンに合わせかつ予測される更新日時の直後に取得するように前記文書アクセス手段を制御することを特徴とする請求項3記載の文書収集装置。

【請求項5】　文書のアクセス時間と前記文書のアクセスに関する平均データ伝送率を記録するデータ伝送率記録手段を更に備え、前記文書アクセス制御手段が平均データ伝送率の大きい文書を優先して取得するように前記文書アクセス手段を制御することにより効率的な収集を可能にしたことを特徴とする請求項2記載の文書収集装置。

【請求項6】　前記データ伝送率記録手段に記録された前記文書のアクセスに関する平均データ伝送率に基づき、前記文書アクセス制御手段が平均データ伝送率の異なる複数の文書へのアクセスを組み合せたアクセスプランを作成し、前記文書アクセス手段が複数の文書を並行にアクセスすることにより、ネットワーク資源から得られる可能な限りのデータ伝送率をより均一かつ効率的に利用できるようにしたことを特徴とする請求項5記載の

2

文書収集装置。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、ネットワーク上に配置された文書群のデータを該ネットワークに接続された特定の計算機内に自動的に収集する文書収集装置に関し、特に文書の取得履歴を記録し、その取得履歴に基づき最新文書の取得を自動的に行なう文書収集装置に関するものである。

【0002】

【従来の技術】従来、ネットワーク上の文書群のデータを自動的に収集する方法として、図11に示すネットワークロボットと呼ばれる文書収集手段がある。このネットワークロボットは、ネットワーク上の文書群の格納位置を与えて置くと、その文書群のデータを自動的に収集してくる機構である。このネットワークロボットの構成は図12のようになる。

【0003】また、特開平6－301577号公報のように、情報源に固有の知識を持ち、その情報源に文書収集装置を派遣し最新の情報を得るものがある。

【0004】

【発明が解決しようとする課題】しかしながら、前記従来のネットワークロボットでは，指定された文書を指定された順序に従ってすべて収集するため、ネットワーク上での文書の更新のパターンによっては、古いデータを収集してしまう、前回収集時以降更新されていないために収集する必要のない文書を収集してしまう、文書へのアクセスに関わるデータ伝送率などの資源を有効に利用できないなどの欠点があった。

【0005】また、情報源に文書収集装置を派遣するものは、派遣された情報源で情報の更新の有無をチェックするため、複数の情報源を考慮した場合の収集の最適化や資源の有効利用が困難であり、文書収集の効率化に限界があった。

【0006】本発明は、前記従来の課題を解決するもので、ネットワーク上にある文書を収集する際に、常に最新の文書情報を自動的にかつ効率良く収集するための文書収集装置を提供することを目的とする。

【0007】

【課題を解決するための手段】この課題を解決するため、本発明は、文書収集をする文書アクセス手段と，それを駆動するアクセス制御手段をもうけ、文書の取得履歴を記録し、その取得履歴に基づき、最新文書の取得を自動的に行なうようにする。

【0008】

【発明の実施の形態】本発明の請求項1記載の発明は、ネットワーク上に分散格納された文書群を自動的に収集する文書収集装置において、文書収集をする文書アクセス手段と，それを駆動するアクセス制御手段をもうけ、前回収集した時点以降に更新された文書のみを収集する

ことを特徴とする文書収集装置であり、最新の文書を効率的に収集保持できるという作用を有する。

【０００９】本発明の請求項２記載の発明は、文書が格納されているネットワーク上の位置と文書の対応を管理する格納位置管理手段と、ネットワーク上の文書の内容を読み込む文書アクセス手段と、読み込んだ文書の内容を記憶する文書格納手段と、時間を計測する時間計測手段と、前記文書を読み込んだ日時を前記時間計測手段からの時刻で記録する取得履歴記録手段と、前記文書を最後に取得した取得日時とネットワーク上での前記文書の更新日時を比較して前回収集した時点以降に更新された文書のみを収集するために前記文書アクセス手段を制御するアクセス制御手段から構成されることを特徴とする文書収集装置であり、請求項１記載の発明と同様な作用を有する。

【００１０】本発明の請求項３記載の発明は、文書とともに得られた更新日時を更新履歴として記録する更新履歴記録手段を更に備え、前記更新履歴記録手段に記録された文書毎の複数の更新日時の記録により、前記アクセス制御手段が前記文書の平均的な更新周期を計算し前記文書を前記更新周期毎かつ予測される更新日時の直後に取得するように前記時間計測手段を利用して前記文書アクセス手段を制御することを特徴とする請求項２記載の文書収集装置としたものであり、適切な更新時刻を予想して文書収集することができるため、一度に文書収集する場合に比べ負荷の分散をすることができるという作用と、常に最新の文書を収集保持することができるという作用を有する。

【００１１】本発明の請求項４記載の発明は、前記更新履歴記録手段に記録された文書毎の複数の更新日時の記録により更新パターンを抽出する更新パターン抽出手段と、その更新パターンを記録する更新パターン記憶手段とを更に備え、前記文書を前記更新パターンに合わせかつ予測される更新日時の直後に取得するように前記文書アクセス手段を制御することを特徴とする請求項３記載の文書収集装置としたものであり、更新パターンを抽出し、その更新パータン毎に文書を収集することで、更新パターンにあった時だけ文書収集し、平均的に最新の文書収集ができるという作用を有する。

【００１２】本発明の請求項５に記載の発明は、文書のアクセス時間と前記文書のアクセスに関する平均データ伝送率を記録するデータ伝送率記録手段を更に備え、前記文書アクセス制御手段が平均データ伝送率の大きい文書を優先して取得するように前記文書アクセス手段を制御することにより効率的な収集を可能にしたことを特徴とする請求項２記載の文書収集装置としたものであり、平均データ伝送率の小さいいくつかの文書を収集しないことなどにより、効率的な文書収集が可能であるという作用を有する。

【００１３】本発明の請求項６に記載の発明は、前記データ伝送率記録手段に記録された前記文書のアクセスに関する平均データ伝送率に基づき、前記文書アクセス制御手段が平均データ伝送率の異なる複数の文書へのアクセスを組み合せたアクセスプランを作成し、前記文書アクセス手段が複数の文書を並行にアクセスすることにより、ネットワーク資源から得られる可能な限りのデータ伝送率をより均一かつ効率的に利用できるようにしたことを特徴とする請求項５記載の文書収集装置としたものであり、前記ネットワーク資源から得られる可能な限りのデータ伝送率をより均一かつ効率的に利用できるという作用を有する。

【００１４】以下、本発明の実施の形態について、図１から図１０を用いて説明する。

【００１５】（第１の実施の形態）本発明の第１の実施の形態について、図１を参照しながら説明する。図１は本発明の第１の実施の形態における文書収集装置の構成を示す概念図である。図１において、文書収集装置は、文書が格納されているネットワーク上の位置と文書の対応を管理する格納位置管理手段101と、ネットワーク上の文書の内容を読み込む文書アクセス手段102と、読み込んだ文書の内容を記憶する文書格納手段103と、前記文書を読み込んだ日時を後記する時間計測手段106からの時刻で記録する取得履歴記録手段104と、前記文書を最後に取得した取得日時とネットワーク上での前記文書の更新日時を比較して前回収集した時点以降に更新された文書のみを収集するために前記文書アクセス手段102を制御するアクセス制御手段105と、時間を計測する時間計測手段106とから構成されている。

【００１６】以上のように構成された文書収集装置について、以下その動作を説明する。図２は文書収集装置の動作手順を示す。

【００１７】まず、ステップ202において、アクセス制御手段105が文書アクセス手段102を起動する。ところで、格納位置管理手段101には文書アクセス手段102がアクセスするネットワークでの文書の位置と文書名が格納されている。

【００１８】例えば、文書名、文書位置の順に、
”文書A http://a/b/c1.html”
のような記述が複数格納されているとする。

【００１９】また、履歴記録手段104は、前回に文書を取得した文書名とその取得日時が、例えば、
”文書A Got 1996-03-18-10:55:30”
のように記録されている。

【００２０】ステップ203では、格納位置管理手段101の文書名、文書位置の組を全てチェックしたか判定する。まだ、何もチェックしていないので、ステップ204に進む。

【００２１】ステップ204において、文書アクセス手段102は、格納位置管理手段101から文書名、文書位置を一つ得る。また、履歴記録手段104から文書名にマッチす

る取得時間を得て、それらを使用してネットワークから目的の文書を得る。例えば、

" 文書A http://a/b/c1.html" と" 文書A Got 1996-03-18-10:55:30"

から文書格納手段103に、前回取得した日時以降に更新された文書のみを得る。

【００２２】また、履歴記録手段104に時間計測手段106から現在の時刻を得て、

" 文書A Got 1996-04-18-10:55:30"

のようにその内容を更新する。

【００２３】この動作を、ステップ203において、格納位置管理手段101内の文書名、文書位置の組を全てチェックし終わるまで繰り返す。

【００２４】以上のように、第１の実施の形態によれば、最新の文書収集を自動的に行なってこれを保持することができるという効果を有する。

【００２５】（第２の実施の形態）本発明の第２の実施の形態について、図３を参照しながら説明する。図３は本発明の第２の実施の形態における文書収集装置の構成を示す概念図である。図３において、文書収集装置は、文書が格納されているネットワーク上の位置と文書の対応を管理する格納位置管理手段301と、ネットワーク上の文書の内容を読み込む文書アクセス手段302と、読み込んだ文書の内容を記憶する文書格納手段303と、前記文書を読み込んだ日時を後記する時間計測手段306からの時刻で記録する取得履歴記録手段304と、前記文書を最後に取得した取得日時とネットワーク上での前記文書の更新日時を比較して前回収集した時点以降に更新された文書のみを収集するために前記文書アクセス手段302を制御するアクセス制御手段305と、時間を計測する時間計測手段306と、文書が更新された日時を記録する更新履歴記録手段307とから構成されている。

【００２６】以上のように構成された文書収集装置について、以下その動作を説明する。図４は文書収集装置の動作手順を示す。

【００２７】まず、更新履歴記録手段307は、文書名とその文書が更新された日時の履歴を、例えば、

" 文書A Modified 1996-03-01-10:01:00
1996-04-01-10:01:00
1996-05-01-10:01:00"

のように記録しているものとする。

【００２８】そこで、ステップ402において、アクセス制御手段305は、更新履歴記録手段307内の更新データから、平均的な更新周期を計算し、最後に更新された日時にその平均周期を加えた時刻（予想更新時刻）を計算する。

【００２９】次にステップ403において、時間計測手段306から現在の時刻を得て、予想更新時刻を経過したか判定する。

【００３０】そして、ステップ404において、予想更新

時刻を経過したら、その文書を文書アクセス手段302を起動することでアクセスする。例えば、この例では、文書Aは、平均更新周期は約１カ月なので、最後の

" Modified 1996-05-01-10:02:00"

に１カ月を加えた

" Modified 1996-06-01-10:02:00"

の時刻（予想更新時刻）を経過したら、目的の文書Aに文書アクセス手段302を起動してアクセスする。

【００３１】ここで、格納位置管理手段301には、文書アクセス手段302がアクセスするネットワークでの文書の位置と文書名を格納されているものとする。例えば、文書名、文書位置の順に、

" 文書A http://a/b/c1.html"

のような記述が複数格納されているとする。

【００３２】また、取得履歴記録手段304には、前回に文書を取得した文書名とその取得日時が、

" 文書A Got 1996-03-18-10:55:30"

のように記録されているものとする。

【００３３】ステップ405において、文書アクセス手段302が格納位置管理手段301からアクセス制御手段305で指定された文書名と文書位置を一つ得る.取得履歴記録手段304から文書名にマッチする取得時間を得て、それらを使用してネットワークから目的の文書を得る。例えば、

" 文書A http://a/b/c1.html" と" Got 1996-03-18-10:55:30"

から文書格納手段303に前回取得した日時以降に更新された文書のみを得る。

【００３４】また、取得履歴記録手段304に時間計測手段306から現在の時刻を得て

" 文書A Got 1996-06-01-10:03:30"

のようにその内容を更新する。また、更新履歴記録手段307に取得文書の更新日時を追加する。そして、ステップ402から繰り返す。

【００３５】以上のように、第２の実施の形態によれば、常に最新の文書収集を自動的に行なってこれを保持することができるという効果を有する。

【００３６】（第３の実施の形態）本発明の第３の実施の形態について、図５を参照しながら説明する。図５は本発明の第３の実施の形態における文書収集装置の構成を示す概念図である。図５において、文書収集装置は、文書が格納されているネットワーク上の位置と文書の対応を管理する格納位置管理手段501と、ネットワーク上の文書の内容を読み込む文書アクセス手段502と、読み込んだ文書の内容を記憶する文書格納手段503と、前記文書を読み込んだ日時を時間計測手段506からの時刻で記録する取得履歴記録手段504と、前記文書を最後に取得した取得日時とネットワーク上での該文書の更新日時を比較して前回収集した時点以降に更新された文書のみを収集するために文書アクセス手段502を制御するアク

セス制御手段505と、時間を計測する時間計測手段506と、文書が更新された日時を記録する更新履歴記録手段507と、更新履歴記録手段507から各文書の更新パターンを抽出する更新パターン抽出手段508と、更新パターン抽出手段508が抽出した更新パターンを記録する更新パターン記録手段509とから構成されている。

【００３７】以上のように構成された文書収集装置について、以下その動作を説明する。図６は文書収集装置の動作手順を示す。

【００３８】まず、更新履歴記録手段507は、文書名とその文書が更新された日時の履歴を、例えば、

" 文書A Modified 1996-03-01-10:01:00
1996-04-01-10:01:00
1996-05-01-10:01:00"

のように記録しているものとする。

【００３９】そしてステップ602において、更新パターン抽出手段508は、更新履歴記録手段507内の文書履歴から、各文書の平均的な更新周期を計算し、同じ更新周期に対応する各文書を更新パターンとして、更新パターン記録手段509に記録する。

【００４０】ところで、更新パターン記録手段509は、更新パターンと文書名を、

" １年毎:文書B,文書C" と " １ヶ月:文書A"

のように記録しているものとする。これは、文書B、文書Cは１年毎に、文書Aは１カ月毎に更新していることを表す。

【００４１】ステップ603において、アクセス制御手段505は、更新パターン記録手段509から更新周期が短いものから順に更新周期とこの更新周期に属する文書名を得る。さらに更新履歴記録手段507から一致する文書名で、最後に更新された日時に、この更新周期を加えた日時（予想更新時刻）を計算する。これを全ての更新時刻について行ない、得られた予想更新時刻を時刻の早いものから順にソートする。

【００４２】ステップ604において、時間計測手段506から得た現在時刻は、ソートされた予想更新時刻にあるか判定する。

【００４３】ステップ605において、予想更新時刻を経過したものがあれば、その文書を文書アクセス手段502を起動することでアクセスする。例えば、この例では、文書Aは平均更新周期は約１ヶ月なので、最後の

" Modified 1996-05-01-10:02:00"

に１ヶ月を加えた

" Modified 1996-06-01-10:02:00"

の時刻（予想更新時刻）を経過したら、目的の文書Aに文書アクセス手段502を起動してアクセスする。

【００４４】ここで、格納位置管理手段501には、文書アクセス手段502がアクセスするネットワークでの文書の位置と文書名を格納されているものとする。例えば、文書名、文書位置の順に、

" 文書A http://a/b/c1.html"

のような記述が複数格納されているとする。

【００４５】また、取得履歴記録手段504には、前回に文書を取得した文書名とその取得日時が、

" 文書A Got 1996-03-18-10:55:30"

のように記録されているものとする。

【００４６】ステップ606において、文書アクセス手段502が格納位置管理手段501からアクセス制御手段505で指定された文書名と一致する文書位置を得る。また、取得履歴記録手段504から文書名にマッチする取得時間を得て、それらを使用してネットワークから目的の文書を得る。例えば、

" 文書A http://a/b/c1.html" と " 文書A Got 1996-03-18-10:55:30"

から文書格納手段503に前回取得した日時以降に更新された文書のみを得る。また、取得履歴記録手段504に時間計測手段506から現在時刻を得て、

" 文書A Got 1996-06-01-10:03:30"

のようにその内容を更新する。また、更新履歴記録手段507に取得文書の更新日時を追加する。そして、ステップ603から繰り返す。

【００４７】以上のように、第３の実施の形態によれば、更新パターンにあった時だけ文書収集を行ない、平均的に最新の文書収集を効率よく自動的に行なってこれを保持することができるという効果を有する。

【００４８】（第４の実施の形態）本発明の第４の実施の形態について、図７を参照しながら説明する。図７は本発明の第４の実施の形態における文書収集装置の構成を示す概念図である。図７において、文書収集装置は、文書が格納されているネットワーク上の位置と文書の対応を管理する格納位置管理手段701と、ネットワーク上の文書の内容を読み込む文書アクセス手段702と、読み込んだ文書の内容を記憶する文書格納手段703と、前記文書を読み込んだ日時を後記する時間計測手段706からの時刻で記録する取得履歴記録手段704と、後記するデータ伝送率記録手段707に記録された前記文書のアクセス時間と平均データ転送率に基づき文書アクセス手段702を制御するアクセス制御手段705と、時間を計測する時間計測手段706と、前記文書を読み込んだ際のアクセス時間と平均データ転送率を記録するデータ伝送率記録手段707とから構成されている。

【００４９】以上のように構成された文書収集装置について、以下その動作を説明する。図８は文書収集装置の動作手順を示す。ここで、データ伝送率記録手段707には、各文書への過去のアクセスに基づき、文書のアクセス時間と該文書のアクセスに関する平均データ伝送率が記録されているものとする。

【００５０】ステップ802において、まず、アクセス制御手段705は、データ伝送率記録手段707に記録されている平均データ伝送率が大きな文書から順にアクセスする

ように、文書アクセス手段702を起動する。この時、時間計測手段706よりアクセス開始時刻を得る。

【００５１】ステップ803において、文書アクセス手段702は格納位置管理手段701から、指定された文書名に一致する文書位置を得る。また、取得履歴記録手段704から、この文書の前回得た日時を得る。そして、この文書をネットワーク上の文書群から指定された文書位置で、前回得た日時以降に更新された文書のみを得て、文書格納手段703に格納する。この時、時間計測手段706から現在の時刻を得て、取得履歴記録手段704に記録する。

【００５２】ステップ804において、アクセス制御手段705は、時間計測手段706からアクセス終了時刻を得て、アクセス開始時刻からの経過時間を計算する。また、文書格納手段703から取得文書のサイズを得る。これらの文書サイズと経過時間から、データ伝送率を計算し、データ伝送率記録手段707に記録する。

【００５３】ステップ805において、データ伝送率記録手段707内の全ての文書をチェックしたか判定する。ステップ805でチェックしていなければ、ステップ802から繰り返す。ステップ805でチェックしていれば、ステップ806に進み終了する。

【００５４】以上のように、第４の実施の形態によれば、アクセス制御手段は、平均データ伝送率の順に各文書を読み込むよう文書アクセス手段を制御するので、これにより、収集すべき全文書中の大部分をより早く収集できる。また，平均データ伝送率の小さいいくつかの文書を収集しないことにより、効率的な文書収集が可能である。

【００５５】（第５の実施の形態）本発明の第５の実施の形態について、図９を用いて説明する。本発明の第５の実施の形態における文書収集装置の構成は図７に示される前記第４の実施の形態の文書収集装置の構成と変わりがない。したがって、文書収集装置の動作について説明する。図９は文書収集装置の動作手順を示す。ここでデータ伝送率記録手段707には、各文書への過去のアクセスに基づき、文書のアクセス時間と該文書のアクセスに関する平均データ伝送率が記録されているものとする。

【００５６】ステップ902において、まず、アクセス制御手段705は、データ伝送率記録手段707に記録されている平均データ伝送率に基づき、アクセスプランを作成する。アクセスプランとは平均データ伝送率が異なる文書へのアクセスを組み合わせることにより、ネットワーク資源から得られる可能な限りのデータ伝送率をより均一かつ効率的に利用するものである。アクセスプランの作成方法には、いわゆる組み合せ最適化理論に基づく各種の方法がありうるが、すべての場合に最良解を出す方法はない。また、実際に文書を収集すると、最初に予測したアクセス時間と平均データ伝送率とは一般に差異を生ずるため、段階的にアクセスプランを修正して行く必要

がある。

【００５７】図１０を用いてアクセスプラン作成方法の一例を説明する。このアクセスプランの目的は、与えられたデータ伝送率という資源を最大限利用して、全文書を最短の時間で収集することにある。

【００５８】まず、予想されるアクセス時間が最長の文書にアクセスする。同時に、残ったデータ伝送率を超えて最も近い平均データ伝送率を持つ文書にアクセスする。それがなければ、平均データ伝送率が最大の文書にアクセスし、再び残ったデータ伝送率に対して同様の文書を探す。利用できるデータ伝送率を超えると、文書の実際のデータ伝送率は少し低下するが、ネットワークが提供するデータ伝送率という資源は最大限利用される。

【００５９】一つの文書の収集が終了した時点で、まだ収集していない文書中から、再び予想されるアクセス時間が最長の文書にアクセスし、残ったデータ伝送率を超えて最も近い平均データ伝送率を持つ文書にアクセスする。すべての文書を収集するまで、同様の手順を続ける。

【００６０】図１０の例では文書Aから文書Fまでの６つの文書を収集している。文書Bと文書Eの実際のデータ伝送率は、利用できるデータ伝送率の上限に達しているために少し低下している。

【００６１】このように作成されたアクセスプランに基づき，ステップ903において、アクセス制御手段705は、１つ以上の文書アクセス手段702を文書名を指定して起動する。すなわち並列にアクセス手段702を起動し、文書収集を並列に行なう。この時、時間計測手段706から現在の時刻を得てアクセス開始時間を得る。

【００６２】ステップ904において、文書アクセス手段702は格納位置管理手段701から、指定された文書名に一致する文書位置を得る。また、取得履歴記録手段704から、この文書を前回得た日時を得る。そして、この文書をネットワーク上の文書群から指定された文書位置で、前回得た日時以降に更新された文書のみを得て、文書格納手段703に格納する。この時、時間計測手段706から現在の時刻を得て、取得履歴記録手段704に記録する。

【００６３】ステップ905において、アクセス制御手段705は、時間計測手段706からアクセス終了時刻を得て、アクセス開始時刻からの経過時間を計算する。また、文書格納手段703から取得文書のサイズを得る。これらの文書サイズと経過時間から、データ伝送率を計算し、データ伝送率記録手段707に記録する。

【００６４】ステップ906において、アクセスプランを全て実行したか判定する。ステップ906でチェックしていなければ、ステップ903から繰り返す。ステップ906でチェックしていれば、ステップ907に進み終了する。

【００６５】以上のように、第５の実施の形態によれば、アクセス制御手段は、アクセスプランに基づいて各文書を読み込んで行くように文書アクセス手段を制御す

るので、これにより、収集すべき全文書中をより早く収集でき、効率的な文書収集が可能である。

【００６６】

【発明の効果】以上のように、本発明は、最新の文書収集を自動的に行なうことができる。また、ネットワーク上の文書の次の更新時刻を予測して文書収集をすることができ、常に最新の文書収集を効率的に自動的に行なうことができる。

【００６７】また、ネットワーク上の文書の更新パターンを得て更新パターン毎に文書収集することができるので、さらに効率的に文書収集ができる。

【００６８】また、平均データ伝送率の大きい文書を優先して取得することができるので、平均データ伝送率の小さいいくつかの文書を収集しないことなどにより、効率的な文書収集ができる。

【００６９】また、文書アクセス手段が複数の文書を並行にアクセスすることができる場合に、ネットワーク資源から得られる可能な限りのデータ伝送率を利用することで、効率的な文書収集ができる。

【図面の簡単な説明】

【図１】第１の実施の形態における文書収集装置の構成を示すブロック図、

【図２】第１の実施の形態における文書収集の手順を示すフロー図、

【図３】第２の実施の形態における文書収集装置の構成を示すブロック図、

【図４】第２の実施の形態における文書収集の手順を示すフロー図、

【図５】第３の実施の形態における文書収集装置の構成を示すブロック図 、

【図６】第３の実施の形態における文書収集の手順を示すフロー図、

【図７】第４及び第５の実施の形態における文書収集装置の構成を示すブロック図、

【図８】第４の実施の形態における文書収集の手順を示すフロー図、

【図９】第５の実施の形態における文書収集の手順を示すフロー図、

【図１０】アクセスプラン作成方法の一例を示す概念図、

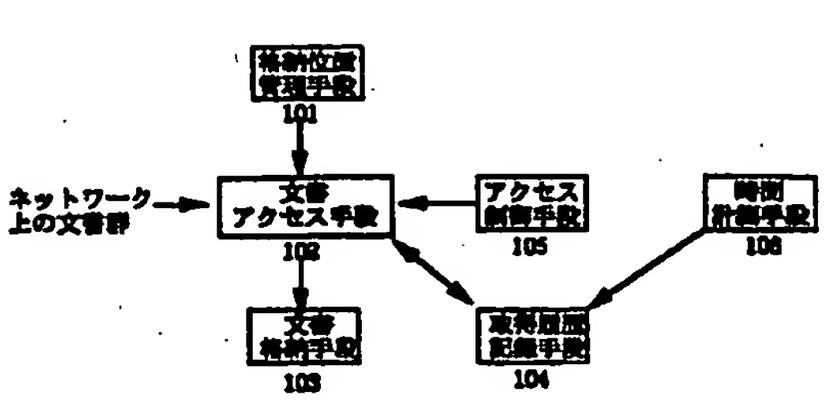
【図１１】ネットワークロボットの使用環境を示す概念図、

【図１２】従来システムの構成を示すブロック図である。

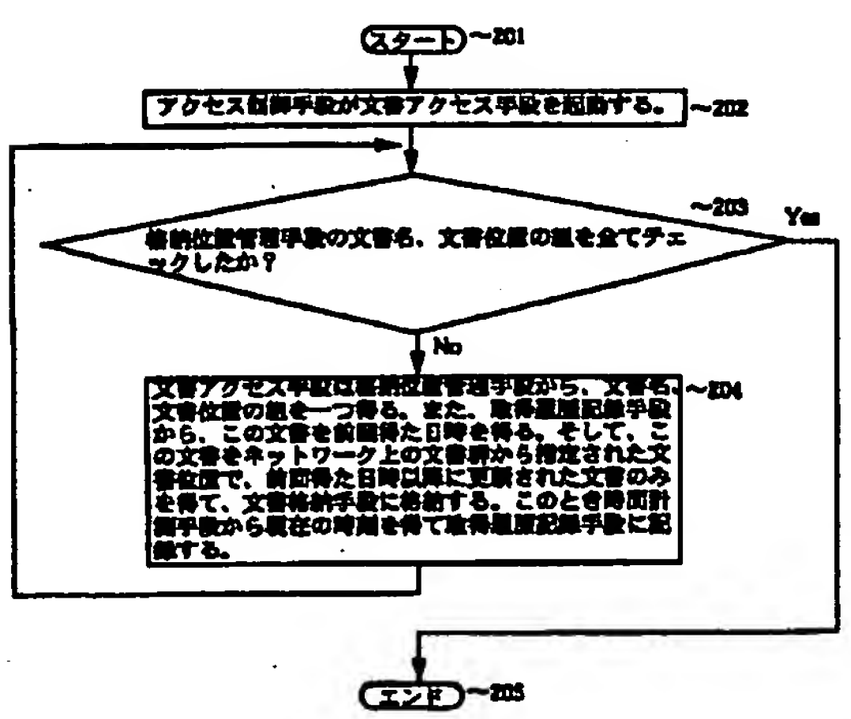
【符号の説明】

101、301、501、701、1201　格納位置管理手段
102、302、502、702、1202　文書アクセス手段
103、303、503、703、1203　文書格納手段
104、304、504、　704　取得履歴記録手段
105、305、505、　705　アクセス制御手段
106、306、506、　706　時間計測手段
307、　507　更新履歴記録手段
508　更新パターン記録手段
509　更新パターン抽出手段
707　データ伝送率記録手段

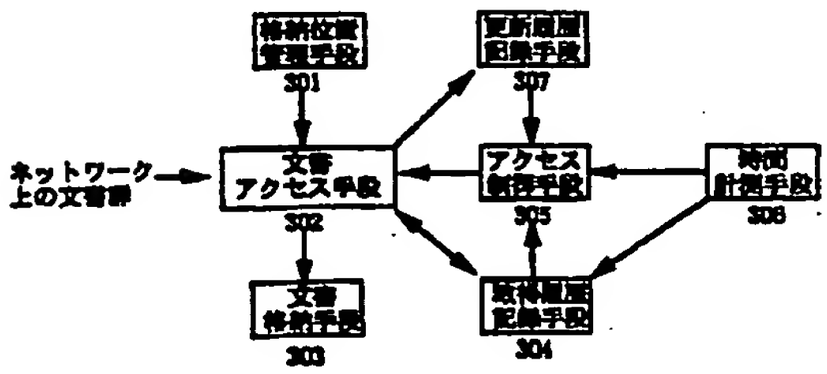
【図１】



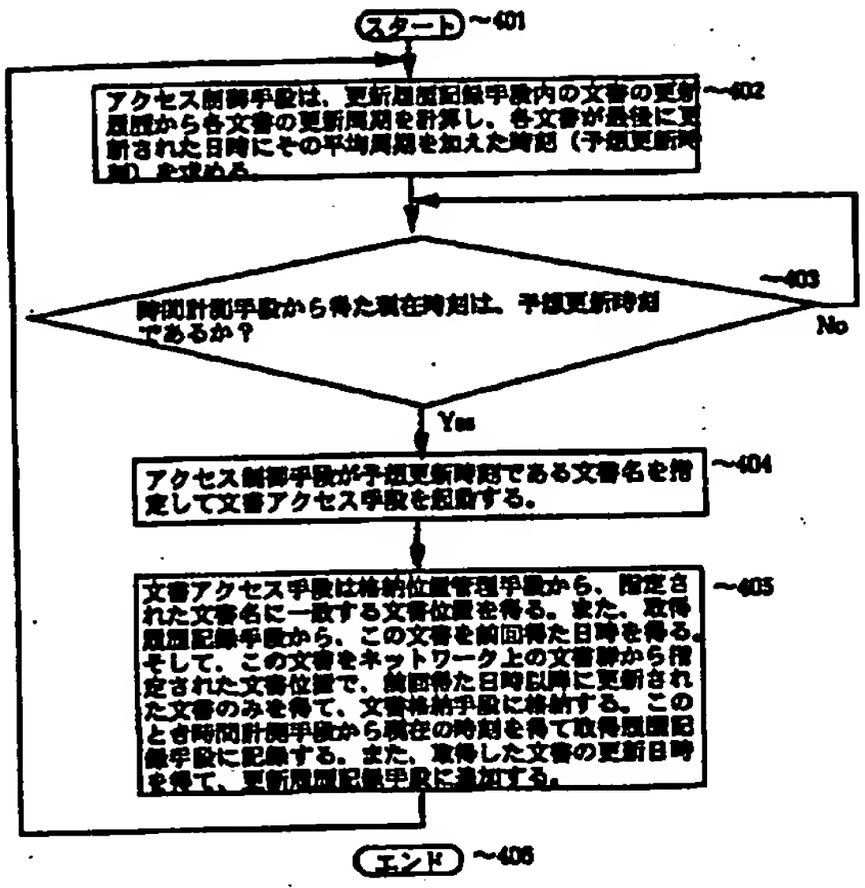
【図２】



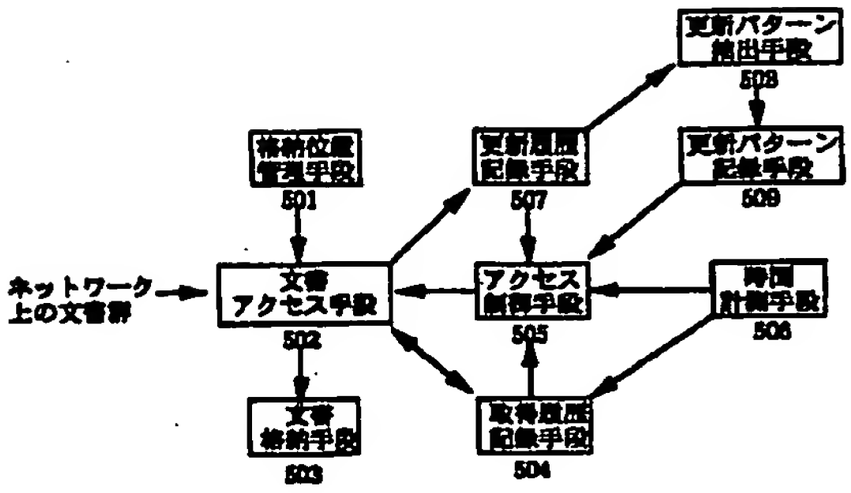
【図３】



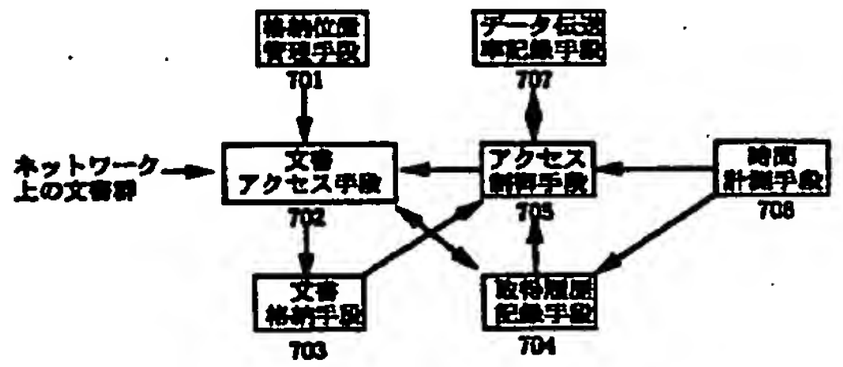
【図4】


スタート
~401
アクセス制御手段は、更新履歴記録手段内の文書の更新履歴から各文書の更新周期を計算し、各文書が最後に更新された日時にその平均周期を加えた時刻（予想更新時刻）を求める。
~402
時間計測手段から得た現在時刻は、予想更新時刻であるか？
~403
No
Yes
アクセス制御手段が予想更新時刻である文書名を指定して文書アクセス手段を起動する。
~404
文書アクセス手段は格納位置管理手段から、指定された文書名に一致する文書位置を得る。また、取得履歴記録手段から、この文書を前回得た日時を得る。そして、この文書をネットワーク上の文書群から指定された文書位置で、前回得た日時以降に更新された文書のみを得て、文書格納手段に格納する。このとき時間計測手段から現在の時刻を得て取得履歴記録手段に記録する。また、取得した文書の更新日時を得て、更新履歴記録手段に追加する。
~405
エンド
~406

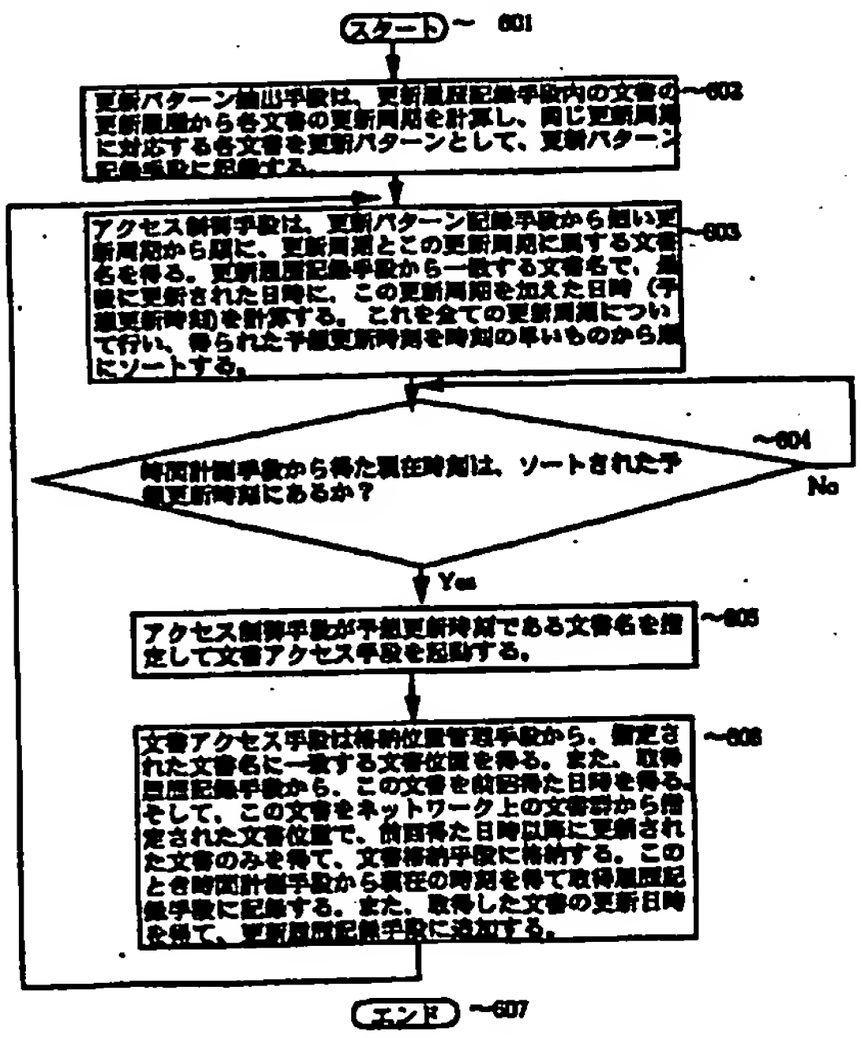
【図5】


更新パターン抽出手段
508
格納位置管理手段
501
更新履歴記録手段
507
更新パターン記録手段
509
ネットワーク上の文書群
文書アクセス手段
502
アクセス制御手段
505
時間計測手段
506
文書格納手段
503
取得履歴記録手段
504

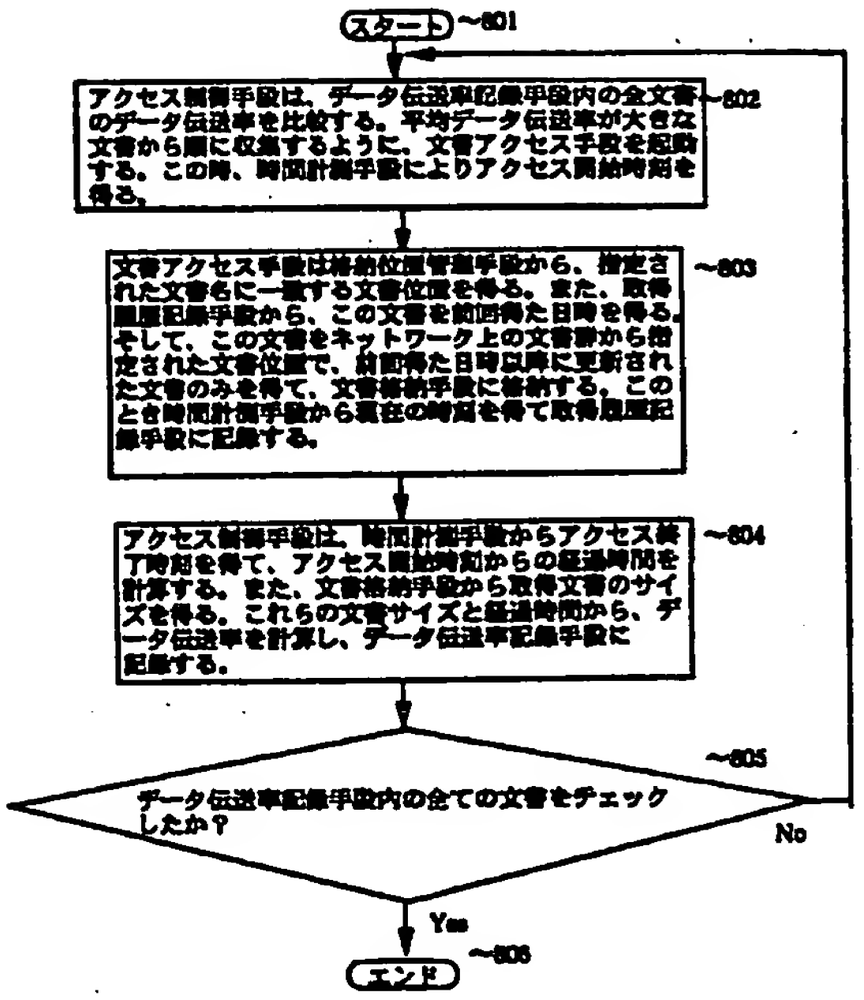
【図7】


格納位置管理手段
701
データ伝送率記録手段
707
ネットワーク上の文書群
文書アクセス手段
702
アクセス制御手段
705
時間計測手段
706
文書格納手段
703
取得履歴記録手段
704

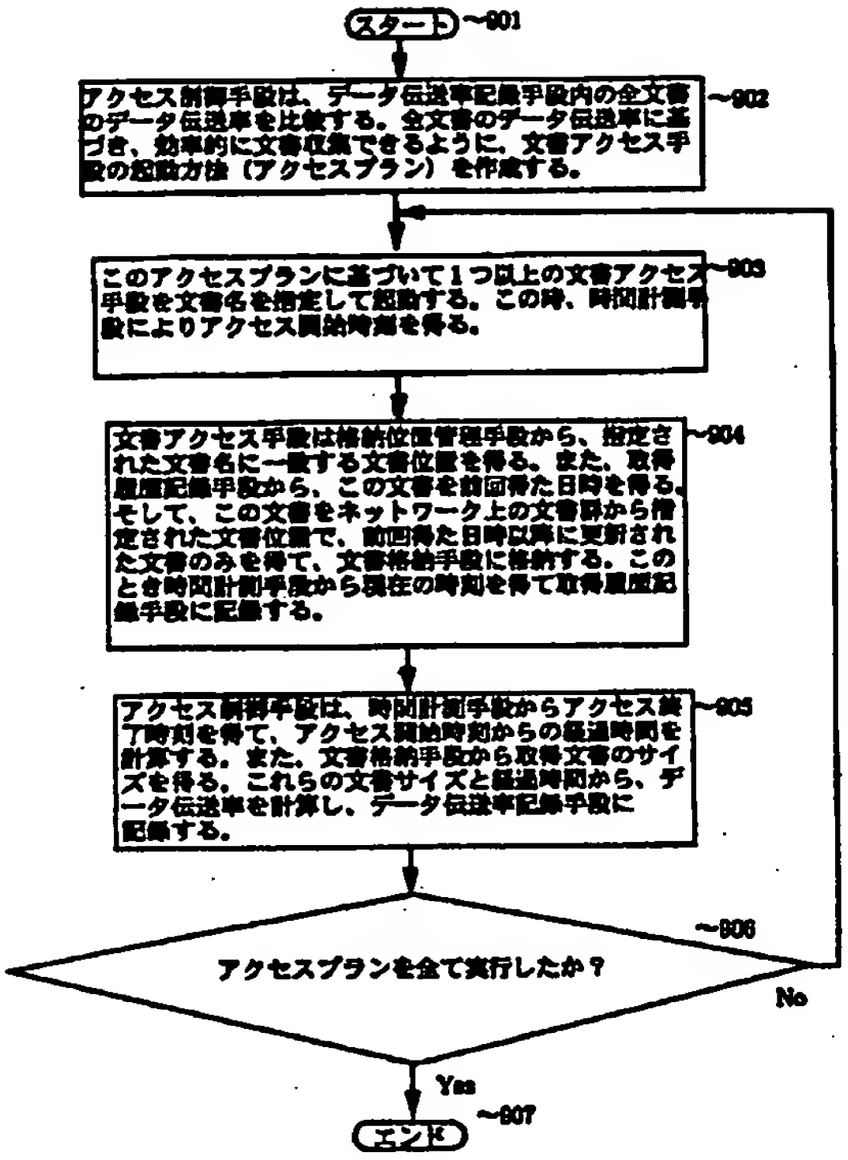
【図6】


スタート
~ 601
更新パターン抽出手段は、更新履歴記録手段内の文書の更新履歴から各文書の更新周期を計算し、同じ更新周期に対応する各文書を更新パターンとして、更新パターン記録手段に記録する。
~602
アクセス制御手段は、更新パターン記録手段から短い更新周期から順に、更新周期とこの更新周期に属する文書名を得る。更新履歴記録手段から一致する文書名で、最後に更新された日時に、この更新周期を加えた日時（予想更新時刻）を計算する。これを全ての更新周期について行い、得られた予想更新時刻を時刻の早いものから順にソートする。
~603
時間計測手段から得た現在時刻は、ソートされた予想更新時刻にあるか？
~604
No
Yes
アクセス制御手段が予想更新時刻である文書名を指定して文書アクセス手段を起動する。
~605
文書アクセス手段は格納位置管理手段から、指定された文書名に一致する文書位置を得る。また、取得履歴記録手段から、この文書を前回得た日時を得る。そして、この文書をネットワーク上の文書群から指定された文書位置で、前回得た日時以降に更新された文書のみを得て、文書格納手段に格納する。このとき時間計測手段から現在の時刻を得て取得履歴記録手段に記録する。また、取得した文書の更新日時を得て、更新履歴記録手段に追加する。
~606
エンド
~607

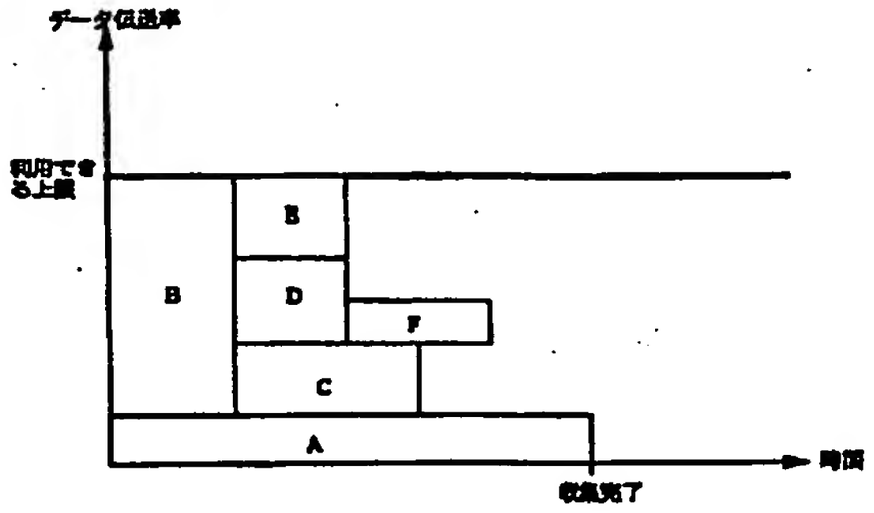
【図8】


スタート
~801
アクセス制御手段は、データ伝送率記録手段内の全文書のデータ伝送率を比較する。平均データ伝送率が大きな文書から順に収集するように、文書アクセス手段を起動する。この時、時間計測手段によりアクセス開始時刻を得る。
~802
文書アクセス手段は格納位置管理手段から、指定された文書名に一致する文書位置を得る。また、取得履歴記録手段から、この文書を前回得た日時を得る。そして、この文書をネットワーク上の文書群から指定された文書位置で、前回得た日時以降に更新された文書のみを得て、文書格納手段に格納する。このとき時間計測手段から現在の時刻を得て取得履歴記録手段に記録する。
~803
アクセス制御手段は、時間計測手段からアクセス終了時刻を得て、アクセス開始時刻からの経過時間を計算する。また、文書格納手段から取得文書のサイズを得る。これらの文書サイズと経過時間から、データ伝送率を計算し、データ伝送率記録手段に記録する。
~804
データ伝送率記録手段内の全ての文書をチェックしたか？
~805
No
Yes
エンド
~806

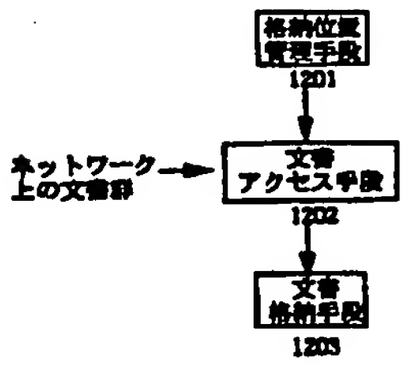
【図９】



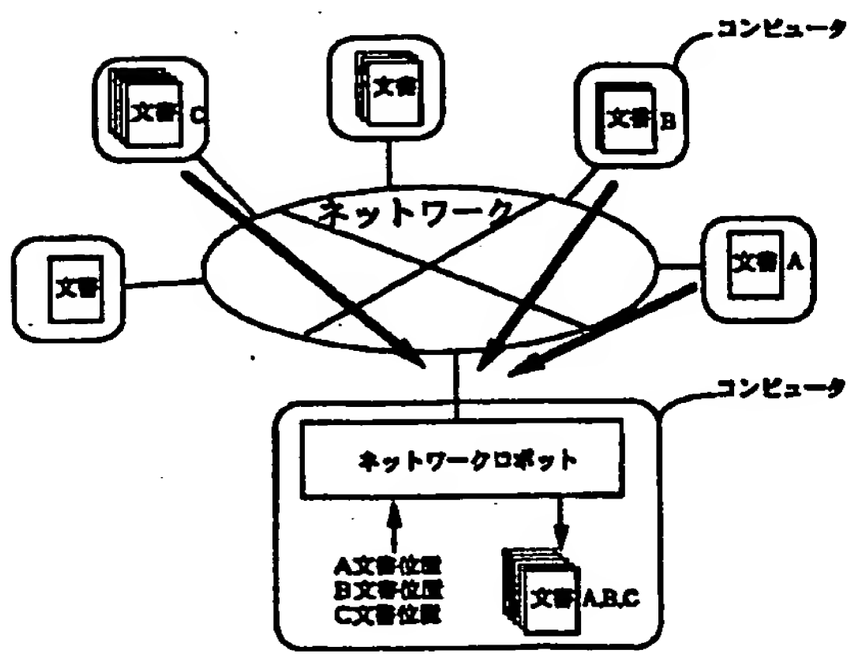
【図１０】



【図１２】



【図１１】



フロントページの続き

(72)発明者　石川　幹人
大阪府門真市大字門真1006番地　松下電器
産業株式会社内